Practical-1

Machine learning basics:

In this lab, we will go through the basics of machine learning. The student needs to make a soft copy note on the following topics:

Topics:

1. **What is Machine learning**

   Machine Learning is a branch of AI and Computer Science which uses data and Algorithms to make machines learn like humans. It is used by many companies for various purposes. For Example, Recommendation, Data Mining, Self-driving cars, etc.

2. **Steps in collection of data**

   1. Identify Opportunities for data collection
   2. Select Opportunities and set goals
   3. Create a plan and set methods for data collection
   4. Validate your systems for measurement
   5. Collect Data
   6. Analyze data
   7. Act based on the data

3. **Steps in importing the data in python (Through: csv, json, and other data formats)**

   CSV: -

   Import pandas as pd

   Dt = pd.read_csv(“file_path.csv”)

   JSON: - Import json F =

   open(data.json) data =

   json.load(f) for i in

   data['emp_details']:

   print(i)

f.close()

Text File: -

Import pandas as pd

Txt = pd.read_table("file_path.txt")

4. Preprocessing

Data preprocessing is the process of converting raw data to clean rata using various techniques.

a) Remove Outliers

An outliner is a data-item that deviates from rest of the data. It can be caused by measurement or execution errors. The removal process of an outliner is same as that of removing a data-item from panda's data-frame.

b) Normalize Datasets, Data encoding

Normalization: -

It is the process to convert numerical values of a dataset between 0 and 1. Used for finding probability using the normalized data.

Data Encoding: -

It is the process of converting categorical data into numeric representation.

c) Handling Missing Data

It is necessary to fill in missing data values in data sets as most of the machine learning models that you want to use will provide an error if you pass NaN values into it. There are various methods in dealing with missing data.

- Delete the column with missing values.
- Deleting the row with missing value.
- Filling the missing value with mean, median or mode value of the column.

5. Machine Models

a) Types of machine learning models – Supervised learning, Unsupervised learning, reinforcement learning.

1. *Supervised Learning:* In supervised learning, the model learns from labelled training data. The data consists of input features and corresponding output labels. The goal of supervised learning is to train a model that can make accurate predictions or classifications when given new, unseen data. Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks.

2. *Unsupervised Learning:* In unsupervised learning, the model learns from unlabelled data. The data only consists of input features without any corresponding output

labels. The goal of unsupervised learning is to find hidden patterns, structures, or relationships within the data. This type of learning is often used for clustering, dimensionality reduction, and anomaly detection tasks. Common unsupervised

learning algorithms include k-means clustering, hierarchical clustering, principal component analysis (PCA), and generative adversarial networks (GANs).

3. *Reinforcement Learning:* Reinforcement learning involves training an agent to interact with an environment and learn from feedback in the form of rewards or penalties. The agent learns through trial and error to maximize the cumulative reward. The model learns by taking actions in the environment, receiving feedback, and adjusting its behaviour based on the feedback. Reinforcement learning is commonly used in scenarios where an agent needs to learn how to make a sequence of decisions or actions, such as game playing, robotics, and autonomous driving.

b) Parameters of machine learning model (Learning rate, regularization, etc.)

The model has parameters that are learned during the training process. These parameters define the relationships between the input features and the output labels or predictions. The learning rate is a hyperparameter that controls the step size at which a machine learning algorithm updates the model's parameters during the optimization process. Regularization is a technique used to prevent overfitting in machine learning models. It adds a penalty term to the loss function, discouraging the model from becoming overly complex and fitting the noise in the training data.

6. Test-train data split: using constant ration, k-fold cross validation

Test-Train Data Split and K-Fold Cross-Validation are two common techniques used to evaluate the performance of machine learning models and assess their generalization ability. The test-train data split using constant ration is a simple and widely used method to evaluate machine learning models. It involves dividing the available dataset into two separate sets: the training set and the test set. The test-train split is typically done using a constant ratio, such as 70%-30%, 80%-20%, or 90%-10%, depending on the size of the dataset.

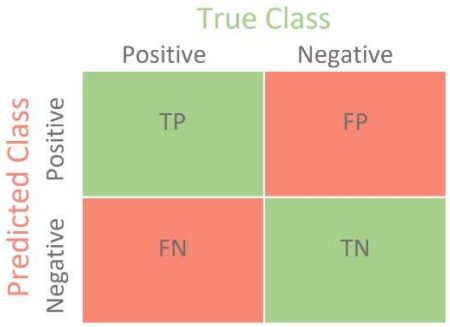
K-Fold Cross-Validation is a resampling technique that helps provide a more robust estimate of a model's performance by repeatedly dividing the dataset into K subsets or "folds." The model is trained and evaluated K times, each time using a different fold as the test set and the remaining folds as the training set.

7. Output Inference

Output interference describes the phenomenon where accuracy decrease over the course of an episodic memory test. Output inference in cued recall takes the form of a decrease in correct and intrusion responses and an increase in failures to response across the test.

8. Validation: different metrics – Confusion Matrix, Precision, Recall, F1-score

Confusion Matrix: - A confusion matrix is a table that summarizes the model's predictions on a classification problem. It shows the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Each entry in the matrix represents the count of instances falling into a specific category.



Precision: - Precision measures the accuracy of positive predictions made by the model. It is defined as the ratio of true positives to the total predicted positives (true positives + false positives). A high precision value indicates that the model makes fewer false positive predictions.

Precision = TP / (TP + FP)

Recall: - Recall measures the ability of the model to identify all positive instances correctly. It is defined as the ratio of true positives to the total actual positives (true positives + false negatives). A high recall value indicates that the model captures a large proportion of positive instances.

Recall = TP / (TP + FN)

F1-score: - The F1-score is the harmonic mean of precision and recall. It provides a single score that balances both metrics. F1-score is useful when you need to consider both precision and recall simultaneously. It ranges between 0 and 1, where 1 is the best possible F1-score.

F1-score = 2 * (Precision * Recall) / (Precision + Recall)